Easy Computation of Bayes Factors and Normalizing Constants for Mixture Models via Mixture Importance Sampling

Mary J. Emond, Adrian E. Raftery and Russell J. Steele
University of Washington [1]

Technical Report no. 398
Department of Statistics
University of Washington

July 1, 2001

[1] Mary J. Emond is Research Assistant Professor of Biostatistics, University of Washington, Box 357237, Seattle, WA 98195-7237; E-mail: `emond@u.washington.edu`. Adrian E. Raftery is Professor of Statistics and Sociology, and Russell J. Steele is Graduate Research Assistant, both at the Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322. E-mail: `raftery/steele@stat.washington.edu`; Web: `www.stat.washington.edu/raftery` and `www.stat.washington.edu/steele`. Emond's research was supported by National Institute of Health Grant 1R29CA77607, and the research of Raftery and Steele was supported by Office of Naval Research Grants N00014-96-1-0192 and N00014-96-1-0330. The authors are grateful to Manisha Desai and Matthew Stephens for helpful discussions.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
01 JUL 2001

2. REPORT TYPE

3. DATES COVERED
00-07-2001 to 00-07-2001

4. TITLE AND SUBTITLE
Easy Computation of Bayes Factors and Normalizing Constants for Mixture Models via Mixture Importance Sampling

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
unclassified	**unclassified**	**unclassified**		**37**	

Abstract

We propose a method for approximating integrated likelihoods, or posterior normalizing constants, in finite mixture models, for which analytic approximations such as the Laplace method are invalid. Integrated likelihoods are key components of Bayes factors and of the posterior model probabilities used in Bayesian model averaging. The method starts by formulating the model in terms of the unobserved group memberships, $\mathbf{Z}$, and making these, rather than the model parameters, the variables of integration. The integral is then evaluated using importance sampling over the $\mathbf{Z}$. The tricky part is choosing the importance sampling function, and we study the use of mixtures as importance sampling functions. We propose two forms of this: defensive mixture importance sampling (DMIS), and Z–distance importance sampling. We choose the parameters of the mixture adaptively, and we show how this can be done so as to approximately minimize the variance of the approximation to the integral.

The resulting method is easy to implement, involving only simple multinomial sampling, it is almost as easy for complex mixture models as for simple ones, and it extends easily to more complicated mixture models. The simulated values on which it is based are independent, and so it avoids problems of convergence due to dependence of successive iterates. We also propose a way of dealing with the label-switching problem. The method provides a standard error, and so is to some extent self-monitoring. In simulations based on a problem in molecular biology, the methods performed well. The approach can be applied more generally to models that are simple when written in a complete-data form, i.e. that are amenable to the EM algorithm, such as models for missing data, for censoring or truncated data, random effects or variance components.

Key Words: Allelotype Bayes factor; Bayesian model averaging; Beta-binomial distribution; Defensive mixture importance sampling; Gibbs sampling; Label-switching; Markov chain Monte Carlo; Multimodality.

Contents

List of Tables

1 Introduction

The integrated likelihood, sometimes also called the marginal likelihood, plays an essential role in Bayesian inference and testing, as it is the central component of the Bayes factor for comparing two models, and of the posterior model probability of one model conditional on data and on a set of candidate models. It also plays a role in Bayesian estimation, as the normalizing constant for the posterior distribution. The integrated likelihood of a model is

$$I \equiv \mathrm{pr}(\mathbf{x}) = \int f(\mathbf{x}|\tau)p(\tau)d\tau, \tag{1}$$

where $\mathbf{x}$ denotes the observed data, $f(\mathbf{x}|\tau, M)$ is the likelihood function for the parameter τ under the model, and $p(\tau)$ is the density (or probability mass function) for the prior distribution of τ given the model.

Since the integrated likelihood often is not analytically tractable, a body of literature on the use of numerical methods for its calculation has developed. These are reviewed in Evans and Swartz (1995) and include methods based on quadrature rules, Laplace's method, importance sampling and Markov Chain Monte Carlo (MCMC). Combinations of MCMC with importance sampling and the Laplace method are considered by Rozenkranz and Raftery (1994), Raftery (1996b) and Lewis and Raftery (1997). The Bayesian Information Criterion (BIC) can be used as an asymptotic approximation to the log Bayes factor (Schwarz, 1978; Kass and Wasserman, 1995).

In finite mixture models, however, none of these methods is fully satisfactory. Two features of mixture models make many current methods for approximating the integrated likelihood problematic. The first is that the model is not "regular" for testing and model selection purposes. In regular models, the log-likelihood becomes approximately elliptically contoured when there are enough data, even when the true parameter values correspond to a lower-dimensional submodel that one is trying to test. In this standard situation, for example, the likelihood-ratio test statistic has an approximate asymptotic chi-squared distribution with degrees of freedom equal to the difference in the number of parameters. This does not hold in finite mixture models whenever one estimates a model with G components but the true number of components is smaller, so that the true parameter values lie on the edge of the parameter space (Lindsay 1995).

A second feature is the "label-switching" problem, namely that the likelihood is invariant to relabelling of the mixture components, and so has $G!$ modes of the same height. Additional local modes are often present (Lindsay, 1995; Titterington, Smith and Makov, 1985), especially when more than G components are fitted (Atwood *et al*, 1992).

The Laplace method (e.g. Tierney and Kadane 1986) provides an analytic approximation to the integrated likelihood based on the assumption that the posterior distribution is approximately elliptically contoured (e.g. Raftery 1996a), and when this assumption holds it can provide approximations of remarkable quality (e.g. Tierney and Kadane 1986; Grunwald, Guttorp and Raftery 1993; Lewis and Raftery 1997). However, for mixture models this assumption fails when the model being fit has G components and the actual number of components is smaller (Lindsay 1995), which is a situation of great interest for model comparison and testing. Thus the Laplace method does not work in this situation.

The original justification of the BIC was in terms of the Laplace method, and it provides a good approximation to the integrated likelihood in regular models for a unit information prior on the parameters (Kass and Wasserman 1995; Raftery 1995). This justification does not hold for mixture models, although BIC does provide a consistent estimate of the number of components in the mixture (Leroux 1992; Keribin 1998), it leads to density estimates that are consistent for the true density (Roeder and Wasserman 1997), and it has given good results in a range of applications (e.g. Dasgupta and Raftery 1998; Fraley and Raftery 1998, 2000).

Quadrature methods can be used but they begin to break down for problems with more than 4 parameters (Evans and Swartz 1995), and the number of parameters in mixture models quickly surpasses this as the number of groups and/or the dimension of the data increase. When testing or comparing mixture models, one is typically considering at least some models that have substantial numbers of parameters.

Markov chain Monte Carlo (MCMC) can be used to estimate mixture models, and associated methods can be used to approximate integrated likelihoods (e.g. Chib 1995, Raftery 1996b). However, in addition to the usual problems with MCMC methods (dependent samples, convergence issues, complexity of programming and implementation), in mixture models they can easily fall foul of the label-switching problem (Celeux 1997; Stephens 1997, 2000). For example, Neal (1998) pointed out that Chib's (1995) results for a mixture model were in error for this reason. Assessing the accuracy of the estimated integrated likelihoods is not trivial with MCMC because of the dependence between successive samples.

Reversible jump MCMC methods can be used to estimate Bayes factors and posterior model probabilities for mixture models (Richardson and Green 1997), but they are also prone to label-switching problems, and convergence can be even more of a problem than for regular MCMC. For example, in the rejoinder to the discussion of their paper, Richardson and Green (1997) mentioned that diagnostics indicated that their method may not have converged even after 500,000 iterations for the one-dimensional mixture model of the galaxy data they analyzed. The difficulty of implementing reversible jump MCMC efficiently for mixture models seems to increase with the dimension of the data.

Our goal in this paper is to propose importance sampling methods for integrated likelihoods in mixture models that are easy to implement and that avoid the difficulties we have been discussing. The method consists first of reformulating the model in terms of the unobserved group memberships, **Z**, as is done for example for estimation via the EM algorithm, and then estimating the integrated likelihood using importance sampling on the **Z**. Analytic or quadrature methods are used for integration over τ; this is a "complete-data" problem and as such is often straightforward.

The success of any importance sampling method depends critically on the importance sampling function, and here two methods for creating this function are proposed: (1) defensive mixture importance sampling (Hesterberg 1995), and (2) sampling via perturbation of an initial grouping that has high posterior probability (the "Z-distance method"). The second method appears to be new. In both of these methods, the importance sampling function is itself a mixture.

Our goals for integrated likelihood estimators are accurate estimates of the integral, realistic estimates of precision, practicality (in terms of programming time and computational time), and conceptual simplicity. Simulation studies of easy to modestly difficult integration problems suggest that our strategy achieves all of these goals.

The resulting strategy is easy to implement, involving only simple multinomial sampling. The samples on which it is based are independent, so there are no problems of nonconvergence due to dependence between samples, as with MCMC. Our approach includes a way of dealing with the label-switching problem that can plague MCMC, and also provides estimated standard errors, and so is to an extent self-monitoring.

Our approach does not become much more complicated as the complexity of the mixture model increases. Thus it can be used with almost equal ease for high-dimensional mixture models as for univariate ones, and for complex mixture models as for simple ones. It extends easily to most finite mixture models. It seems to provide quite good approximations to the integrated likelihood, and also reliable estimates of the associated standard error.

In Section 2 we review mixture models and present our importance sampling based estimators. In Section 3 we show how these estimators are applied in the context of multimodality due to the label–switching problem. In Section 4 we present simulation results motivated by an application in molecular genetics, and in Section 5 we discuss advantages, limitations and directions for future research.

2 Mixture Models and Importance Sampling Methods

2.1 Finite Mixture Models

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a realization of a random sample $\mathbf{X}^{(n)} \equiv (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ from a G–component mixture distribution. The corresponding likelihood is

$$\prod_{i=1}^{n} \sum_{j=1}^{G} \pi_j f_j(\mathbf{x}_i|\theta_j) \equiv \prod_{i=1}^{n} f(\mathbf{x}_i|\theta, \pi), \tag{2}$$

where the π_j's are mixing proportions that sum to 1, $\pi = (\pi_1, \ldots, \pi_G)$, and the θ_j's are component–specific parameter vectors with $\theta = (\theta_1', \ldots, \theta_G')'$. Each observation, $\mathbf{x}_i$, arises from one of the G component densities, $f_j, j = 1, \ldots, G$, but the group memberships are unknown. The parameter π_j is the unknown probability of an observation arising from f_j.

To obtain the integrated likelihood, or marginal probability of the data, the joint distribution of $\mathbf{x}$ and $\tau = (\theta, \pi)$ is integrated with respect to the unknown parameters:

$$I(\mathbf{x}) \equiv \int_{\tau} \prod_{i=1}^{n} f(\mathbf{x}_i|\theta, \pi) p(\theta, \pi) d\tau, \tag{3}$$

where $p(\theta, \pi)$ is the prior density for $(\theta, \pi) = \tau$. Analytic integration of (3) is virtually always impossible.

The component membership for $\mathbf{x}_i$ may be thought of as an unobserved random variable. When the component membership is known, the likelihood takes a simpler form. Let $\mathbf{z}_i \equiv (z_{i1}, \ldots, z_{iG})'$ be the vector that indicates component membership for the i^{th} observation such that $z_{ij} = 1$ if $\mathbf{x}_i$ is from component j and 0 otherwise. The $n \times G$ matrix $\mathbf{Z}_n \equiv \{\mathbf{z}_1', \ldots, \mathbf{z}_n'\}'$ gives the component membership for the entire sample, and we let $\mathcal{Z}_i$ and $\mathcal{Z}^{(n)}$ be the random variables corresponding to $\mathbf{z}_i$ and $\mathbf{Z}_n$, respectively. Then

$$\begin{aligned} I(\mathbf{x}) &= \int_{\tau} \prod_{i=1}^{n} \int_{\mathbf{z}_i} f_{X|Z}(\mathbf{x}_i|\mathcal{Z}_i = \mathbf{z}_i, \theta, \pi) f_Z(\mathbf{z}_i|\theta, \pi) d\mathbf{z}_i p(\theta, \pi) d\tau & (4) \\ &= \int_{\mathbf{Z}_n} \int_{\tau} \prod_{i=1}^{n} \prod_{j=1}^{G} (\pi_j f_j(\mathbf{x}_i|\theta))^{z_{ij}} p(\theta, \pi) d\tau d\mathbf{Z}_n. & (5) \end{aligned}$$

Note that we have assumed that $f_Z(z_i|\theta, \pi) = f_Z(z_i|\pi) = \prod_{j=1}^{G} \pi_j^{z_{ij}}$. In the rest of this article, we also make the (usually natural) assumption that θ and π are indepedent *a priori*, so that $p(\theta, \pi) = p(\theta)p(\pi)$ (where it is understood that the $p(\cdot)$'s refer to different functions, depending on the argument).

This formulation greatly simplifies part of the problem, since the inner integral of (5) (an integration with respect to $\tau = (\theta, \pi)$) often can be evaluated analytically, or at least closely approximated via the Laplace method or a similar approach, and may also be more amenable to numerical integration via quadrature. The problem takes the following general form:

$$I = \int L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})d\mathbf{Z}. \tag{6}$$

Here, $L(\mathbf{x}|\mathbf{Z}) \equiv L(\mathbf{x}|\mathcal{Z}^{(n)} = \mathbf{Z}_n) = \int_\theta \prod_{i=1}^n f(\mathbf{x}_i|z_i, \theta)p(\theta)d\theta$, and $p(\mathbf{Z}) = p(\mathcal{Z}^{(n)} = \mathbf{Z}_n) = \int_\pi \prod_{i=1}^n f_Z(z_i|\pi)p(\pi)d\pi$. In (6) and in some of what follows, we suppress $\mathbf{x}$ and n in the context of a fixed realization of the data. For the purposes of this article, we will assume that integration with respect to (θ, π) can be done analytically. Desai (2000) and Desai and Emond (2001) treat cases where numerical methods are needed for integration with respect to (θ, π).

2.2 Importance Sampling

The integral with respect to $\mathbf{Z}$ in (6) is the summation over G^n points, which can be done exactly for very small data sets. Otherwise, importance sampling is a possible approach. The importance sampling estimate is given by

$$\hat{I} = \frac{1}{K}\sum_{k=1}^{K} L(\mathbf{x}|\mathbf{Z}_k)\frac{p(\mathbf{Z}_k)}{h(\mathbf{Z}_k)} \equiv \sum_{k=1}^{K} I_k, \tag{7}$$

where the $\mathbf{Z}_k$'s are sampled (simulated) independently from the density $h(\cdot)$. Importance sampling has classical elegance: $\hat{I}$ converges almost surely to I and $\sqrt{K}(\hat{I} - I)/\sqrt{\text{var}(\hat{I})}$ has a standard normal limiting distribution when $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})/h(\mathbf{Z})$ has finite mean and variance under $h(\cdot)$. The latter result can provide a ready means for assessing the precision of $\hat{I}$ through its standard error, provided that the standard error is an accurate estimate of $\sqrt{\text{var}(\hat{I})}$.

The catch to using (7) is the choice of $h(\cdot)$, since this choice determines how easily the sampling can be done in practice and also determines the precision of $\hat{I}$. Easy choices for $h(\cdot)$ often lead to inefficient estimates. For example, it may be possible to simulate the $\mathbf{Z}$'s easily when the importance sampling function is the prior distribution of $\mathbf{Z}$, i.e. when $h(\mathbf{Z}) = p(\mathbf{Z})$. When the data are reasonably informative, this results in sampling mostly $\mathbf{Z}_k$'s for which $L(\mathbf{x}|\mathbf{Z}_k)$ has no substantial mass, leading to a very inefficient estimator (Raftery, 1996b).

Worse yet, the empirical variance of the I_k's can markedly underestimate the true variance of $\hat{I}$ under these circumstances, leading to false confidence in the estimate. Stephens and Donnelly (2000) provide an extreme example of this phenomenon in practice.

The problem, then, is to find a good choice of $h(\mathbf{Z})$: one that is reasonably easy to sample from, and that also provides good efficiency. The choice of $h(\mathbf{Z})$ that minimizes $\text{var}(\hat{I})$ is $p(\mathbf{Z}|\mathbf{x}) \equiv p(\mathcal{Z}^{(n)} = \mathbf{Z}|\mathbf{X}^{(n)} = \mathbf{x})$ (c.f. Zhang, 1996; Appendix A). This minimum variance is zero, but one needs to know I in order to know $p(\mathbf{Z}|\mathbf{x})$, so sampling directly from $p(\mathbf{Z}|\mathbf{x})$ is not possible. A potential surrogate for $p(\mathbf{Z}|\mathbf{x})$ is $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ where $\hat{\tau}$ is the maximum likelihood estimate (MLE) for τ in (2). We expect this to be a reasonable starting point for constructing $h(\mathbf{Z})$, since $p(\mathbf{Z}|\mathbf{X}^{(n)})/p(\mathbf{Z}|\mathbf{X}^{(n)}, \tau = \hat{\tau})$ converges in probability to a constant for a sequence of $\mathbf{Z}$'s near the mode for $\mathcal{Z}^{(n)}$ under regularity conditions (see Theorem 1). Sampling from $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ has also been suggested by Wei and Tanner (1990) in the different context of sampling from a posterior distribution; they called this method Poor Man's Data Augmentation, or PMDA.

In the present context, sampling directly from $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ has at least three potential pitfalls. The first is that, unless the likelihood is extremely peaked, it tends to miss $\mathbf{Z}$'s corresponding to important values of $L(\mathbf{x}|\mathbf{Z})$, as the sampling can stick to small subregions of the sample space for seemingly reasonable values of K. In this situation, the sample variance of the I_k's tends to underestimate the true variance of $\hat{I}$. The second pitfall is that $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ is a good approximation to $p(\mathbf{Z}|\mathbf{x})$ only near the modal value of $\mathbf{Z}$, and the third is that the approximation, so far, is justified only when the fitted model is correct with no superfluous components included (Theorem 1). In this paper we propose two methods for constructing importance sampling distributions that use $\hat{\tau}$ and/or $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ to create efficient importance sampling distributions while attempting to circumvent these pitfalls.

2.3 Defensive Mixture Importance Sampling

Here we propose the use of importance sampling distributions that are themselves mixtures, with one component equal to the prior distribution so as to ensure coverage of the area in which the posterior density is nonnegligible. We start with distributions of the form

$$h(\mathbf{Z}) = (1 - \delta)g(\mathbf{Z}) + \delta p(\mathbf{Z}), \ \delta \in [0, 1] \tag{8}$$

for calculation of (7), where $g(\mathbf{Z})$ is an intelligently-chosen probability mass function for $\mathbf{Z}$, and $p(\mathbf{Z})$ is, as before, the marginal prior distribution of $\mathbf{Z}$. The idea of using importance sampling distributions that are themselves mixtures was discussed by Geyer (1991), Oh and Berger (1993), West (1993), Givens and Raftery (1996) and Raghavan and Cox (1998) in different contexts. The idea of using mixtures of the form (8) as importance sampling distributions for evaluating integrated likelihoods was mentioned by Newton and Raftery (1994) and discussed by Raftery (1996b). However, the general approach of using mixtures of this form as importance sampling functions was discussed in greater depth by Hesterberg (1995). He called this a "defensive mixture" for the importance sampling distribution, and so we will refer to this approach as defensive mixture importance sampling (DMIS).

If we define the quantity $p(\mathbf{Z}_k)/h(\mathbf{Z}_k)$ to be the "weight" for the k^{th} term in (7), then the defensive mixture estimate has the property that the weight is bounded by $1/\delta$, guaranteeing the applicability of the Gaussian central limit theorem for $\delta \in [0,1]$ whenever it holds for $\delta = 1$, i.e. for sampling solely from the prior. In addition, the asymptotic variance of the defensive mixture estimate can be no more than $1/\delta$ times the variance obtained by taking $h(\mathbf{Z}) = p(\mathbf{Z})$, resulting in a bound on how much worse DMIS can do relative to taking $h(\mathbf{Z}) = p(\mathbf{Z})$.

In order to improve on $p(\mathbf{Z})$, we need to choose $g(\cdot)$ so that $h(\cdot)$ mimics $p(\mathbf{Z}|\mathbf{x})$. Our proposal is to take $g(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ with $\hat{\tau}$ being the MLE for τ. Note that this is just a multinomial distribution with mean $(\hat{\mathbf{z}}_1 \ldots \hat{\mathbf{z}}_n)'$ where the $\hat{\mathbf{z}}_i = E(\mathcal{Z}_\rangle|\mathbf{x}_\rangle, \tau = \Uparrow)$'s are just the estimated values of the missing $\mathbf{z}_i$'s resulting from the E-step of the EM algorithm used in obtaining $\hat{\tau}$ from (2). This proposal is supported by the following theorem, which indicates that $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$ is near $p(\mathbf{Z}|\mathbf{x})$ when $\mathbf{Z}$ is near the mode of $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ and the fitted model is correct. The superscript (n)'s are included in the statement of the theorem to make the dependence on n explicit.

Theorem 1 *Let $X^{(n)}$ be a random sample from the distribution $f_X(\mathbf{x}|\tau)$ where $f_X(\mathbf{x}|\tau)$ has the representation based on an unobserved variable Z:*

$$f_X(\mathbf{x}|\tau) = \int_Z f_{X|Z}(\mathbf{x}|z, \tau) f_Z(\mathbf{z}|\tau) d\mathbf{z}.$$

Let $\mathcal{Z}^{(n)}$ be the unobserved sample from $f_Z(\mathbf{z}|\tau)$. Let $p_n(\mathbf{Z}_n|\mathbf{X}^{(n)})$ be the probability that $\mathcal{Z}^{(n)} = \mathbf{Z}_n$ given $\mathbf{X}^{(n)}$ and let $p_n(\mathbf{Z}|\mathbf{X}^{(n)}, \tau)$ be the probability that $\mathcal{Z}^{(n)} = \mathbf{Z}$ given $(\mathbf{X}^{(n)}, \tau)$. That is,

$$p_n(\mathbf{Z}_n|\mathbf{X}^{(n)}) = \frac{\int \prod_i f_{X|Z}(\mathbf{X}_i|\mathbf{z}_i, \tau) f_Z(\mathbf{z}_i|\tau) p(\tau) d\tau}{I(\mathbf{X}^{(n)})}, \textit{ and}$$

$$p_n(\mathbf{Z}|\mathbf{X}^{(n)}, \tau) = \prod_i \prod_j \left[\frac{\pi_j f_j(\mathbf{X}_i|\theta_j)}{\sum_j \pi_j f_j(\mathbf{X}_i|\theta_j)} \right]^{z_{ij}} = \prod_i \prod_j p(z_{ij}|\mathbf{X}_i, \tau).$$

Define $\hat{z}_{ij}(\mathbf{X}_i, \tau) = E[\mathcal{Z}_{ij}|\mathbf{X}_i, \tau] = p(z_{ij} = 1|\mathbf{X}_i, \tau)$*, and let* $\hat{\tau}_n$ *be the MLE for* τ *in (6) based on* $\mathbf{X}^{(n)}$*. Hence, the* $\hat{z}_{ij}(\mathbf{X}_i, \hat{\tau}_n)$*'s are the expected values of the missing data used in the EM algorithm (Dempster, Laird and Rubin, 1977). Let* $\hat{z}(\mathbf{X}^{(n)})$ *denote the "sample" of* $\hat{z}_{ij}$*'s formed in this manner. Note that* $\hat{z}_{ij}$ *is not restricted to* $\{0,1\}$*. Define*

$$\tilde{J}(\tau) \equiv -E\left[\frac{\delta^2}{\delta\tau^2}\log f_{X,Z}(\mathbf{X}_i, \hat{z}(\mathbf{X}_i, \tau_0)|\tau)\right] \text{ and },$$

$$J(\tau) = -E\left[\frac{\delta^2}{\delta\tau^2}\log f_X(\mathbf{X}_i|\tau)\right],$$

where τ_0 *represents the true value of* τ *and expectations are taken under the true value. We assume that* $\tilde{J}(\tau_0)$ *and* $J(\tau_0)$ *exist and are positive definite. Assume that conditions (1) – (5) in Appendix A hold. Then*

$$\frac{p_n(\hat{z}(\mathbf{X}^{(n)})|\mathbf{X}^{(n)})}{p_n(\hat{z}(\mathbf{X}^{(n)})|\mathbf{X}^{(n)}, \tau = \hat{\tau}_n)} \xrightarrow{p} \frac{|\tilde{J}(\tau_0)|^{1/2}}{|J(\tau_0)|^{1/2}}. \tag{9}$$

The proof is given in Appendix A.

Remarks: 1. The amount of unknown information in this problem grows in proportion to n. The theorem provides a measure of closeness between $p(\mathbf{Z}_n|\mathbf{x}, \tau = \hat{\tau})$ and $p(\mathbf{Z}_n|\mathbf{x})$ only for a single sequence of $\mathbf{Z}_n$'s. The approximation of $p(\mathbf{Z}_n|\mathbf{x})$ by $p(\mathbf{Z}_n|\mathbf{x}, \tau = \hat{\tau})$ may not be good at points far from $\hat{z}(\mathbf{X}^{(n)})$. This is to be expected, since the data do not provide information about the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$.

2. Condition 2 requires that $\hat{\tau}_n$ converges to a single point. Hence, technically, it may be necessary to restrict the integration to a subset of the parameter space so that $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ is unimodal over this subset. Allowing $\hat{\tau}_n$ to converge over a set of G points still leads to a finite limit, however (see Feng and McCulloch, 1996).

3. Condition (3) in Appendix A and the assumption of positive definiteness for $\tilde{J}(\tau_0)$ and $J(\tau_0)$ both require that $\pi_j > 0,\ j = 1, \ldots G$. It is of interest to know whether these conditions may be relaxed, since we wish to have a good estimate of $p(\mathbf{Z}|\mathbf{x})$ even when we have fit a model with the wrong number of non–zero components. In the situation where too many components are fitted, Feng and McCulloch (1996) have shown that consistency of $\hat{\tau}_n$ still holds, so that $p(\hat{z}(\mathbf{X}^{(n)})|\mathbf{X}^{(n)}, \tau = \hat{\tau}_n)$ will have the same limiting value under the over-fitted model as under the correct model. However, the remainder term in expansion (26) does not

converge to zero in this situation, and $|\tilde{J}(\tau_0)|$ and $|J(\tau_0)|$ are both zero. In this situation, we conjecture that the ratio on the LHS of (9) is $O_p(n^{d/2})$, where d is the number of zero components in the over–fitted model. □

In order to apply the proposed DMIS method, one needs to choose δ. Hesterberg (1995) found that his simulation results were not sensitive to the choice of δ within a range from 0.1 to 0.5. We found that δ's in this range gave reasonably unbiased results in very simple problems, but that the sample variance of the resulting I_k's could vary by a factor of nearly 16 as δ varied, even in these simple problems. Also, the accuracy of the standard errors was sensitive to δ. Hence, a way of choosing δ is needed. The method we have developed chooses δ so as to minimize the variance of $\hat{I}$ under a simplifying approximation, given by the following theorem.

Theorem 2 *Suppose that $L(\mathbf{x}|\mathbf{Z})$ takes on only two distinct values as $\mathbf{Z}$ varies. Define $\mathbf{Z}_M = argmax_{\mathbf{Z}}\{L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})\}$. Let δ_{opt} be the value of δ that minimizes the variance of $\hat{I}$ in (7) over all $h(\mathbf{Z})$ of the form in (8), and let $v_2(\delta) \equiv var(\hat{var}(\hat{I}))$ with $\hat{var}(\hat{I}) = K^{-1}\sum_k (I_k - \bar{I})^2$, the sample variance of I_k's. Then*

$$\delta_{opt} = \frac{Ip(\mathbf{Z}_M|\mathbf{x}, \tau = \hat{\tau}) - L(\mathbf{x}|\mathbf{Z}_M)p(\mathbf{Z}_M)}{Ip(\mathbf{Z}_M|\mathbf{x}, \tau = \hat{\tau}) - Ip(\mathbf{Z}_M)}. \tag{10}$$

Moreover, δ_{opt} is the minimizer of $v_2(\delta)$ over $h(\mathbf{Z})$ of the form in (8).

The proof is given in Appendix B. The proof shows that, regardless of the form of $h(\cdot)$, the most efficient $h(\cdot)$ in this simplified situation satisfies

$$h(\mathbf{Z}_M) = p(\mathbf{Z}_M|\mathbf{x}). \tag{11}$$

That is, we choose our importance sampling distribution to match the optimal importance sampling distribution at its modal value, and (11) provides a potential criterion for choosing $h(\cdot)$ regardless of whether $h(\cdot)$ has the mixture form. For peaked likelihoods, $Ip(\mathbf{Z}_M)$ is small relative to $Ip(\mathbf{Z}_M|\mathbf{x}, \tau = \hat{\tau})$ and can be dropped from the expression.

Equation (10) contains two unknown quantities, $\mathbf{Z}_M$ and I, which must both be estimated. For an initial estimate of $\mathbf{Z}_M$, we set the ij^{th} component of $\hat{\mathbf{Z}}_M$ to be

$$\hat{z}_{ij} = 1[j = \max_k \text{prob}(z_{ik} = 1|\mathbf{x}_i, \tau = \hat{\tau})], \tag{12}$$

where $1[\cdot]$ represents the indicator function. That is, we assign the i^{th} observation to the component with maximum posterior probability conditional on $\mathbf{X} = \mathbf{x}$ and $\tau = \hat{\tau}$. $\hat{\mathbf{Z}}_M$'s obtained in this manner from τ's which are other local maxima of the likelihood equation should be examined as potentially better estimates of $\mathbf{Z}_M$. Also, $\mathbf{Z}$'s close to $\hat{\mathbf{Z}}_M$ can be examined to see whether any has a higher value of $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$, providing a better estimate of $\mathbf{Z}_M$.

The value I in (10) is replaced by $\hat{I}$ estimated using $\delta = 0.5$. This produces a two–step adaptive procedure for estimating I. Even though the justification may initially appear to be overly simplistic, we found that choosing δ according to (10) worked well in simulations, resulting in var$(\hat{I})$ at or very near the empirically determined lower bound, even when the assumption of only two values for $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ was far from the truth. In principle, one could take more than two steps in this procedure and iterate to convergence of $\hat{I}$, but in practice we found little advantage to going beyond two steps in our examples.

The defensive mixture method need not be limited to two components for $h(\cdot)$. Because $p(\mathbf{Z}|\mathbf{x}) = \int p(\mathbf{Z}|\mathbf{x},\tau)p(\mathbf{x}|\tau)p(\tau)d\tau/I$ can be approximated arbitrarily closely by a weighted sum of terms of the form $p(\mathbf{Z}|\mathbf{x},\tau_j)$, one might take

$$g(\mathbf{Z}) = \sum_{t=1}^{T} \delta_t p(\mathbf{Z}|\mathbf{x}, \tau = \tau_t), \tag{13}$$

with $T > 1$ and $\sum_t \delta_t = 1 - \delta$. For example, in a data set where there is significant multimodality (beyond that due to label-switching), the $\hat{\tau}$'s in (13) could correspond to the modes of the likelihood surface. The δ and the δ_t's would be fixed at $1/(T+1)$ to obtain $\hat{I}_0$, an initial estimate of I, and then chosen adaptively in a second step by solving the set of equations

$$h(\mathbf{Z}_t) = \hat{p}(\mathbf{Z}_t|\mathbf{x}),\ t = 1, \ldots, T. \tag{14}$$

Here, the $\mathbf{Z}_t$'s are determined by the $\hat{\tau}_t$'s via (12), and $\hat{p}(\mathbf{Z}_t|\mathbf{x}) = L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})/\hat{I}_0$. This variant of the DMIS method is examined in our simulations with Data Set 5 in Section 4.

2.4 The Z–Distance Method

Our second method is a novel sampling method that directly targets for sampling $\mathbf{Z}$ near $\mathbf{Z}_M$. By "$\mathbf{Z}$ near $\mathbf{Z}_M$" we mean near in the sense that $|L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z}) - L(\mathbf{x}|\mathbf{Z}_M)p(\mathbf{Z}_M)|$ is relatively small, making $\mathbf{Z}$ an important point to sample. For motivation, consider the case where $\max_k p(z_{ik} = 1|\mathbf{x}, \tau = \hat{\tau})$ is not much greater than the $p(z_{ik} = 1|\mathbf{x}, \tau = \hat{\tau})$ for other k's. Then, the point near $\mathbf{Z}_M$ with the i^{th} observation re-assigned to another group will have a value of $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ close to $L(\mathbf{x}|\mathbf{Z}_M)p(\mathbf{Z}_M)$. Our second method, "Z–distance" sampling, attempts to target these points with such re-assignments. Z–distance sampling is done as follows:

1. Divide the n observations into an initial grouping consisting of R groups with $R \in \{1, \ldots, n\}$. Each of the R groups comprises observations that are close to each other in the sense that they have high conditional probability of being in the same group. For example, $\|\hat{z}_{i_1} - \hat{z}_{i_2}\|$ is small for observations i_1 and i_2 in the same group, where $\hat{z}_{i_m}$ is defined in Theorem 1. A non-parsimonious clustering algorithm can be used to form the initial groups, or the initial groups can be those given by $\hat{\mathbf{Z}}_M$ in (12). Let r_j be the number of observations in the j^{th} group.

2. A $\mathbf{Z}_k$ in the importance sampling step is formed by taking the r_j observations in the j^{th} group and redistributing them into the G groups according to a Dirichlet–Multinomial distribution with mean parameter $\mu_j = (\mu_{j1}, \ldots, \mu_{jG})$ and dispersion parameter ω_j that may depend on $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$.

In Appendix C, we review the Dirichlet–Multinomial distribution with the above parametrization and provide details on how to carry out the sampling in practice. The choice of the Dirichlet–Multinomial distribution in step 2 allows for correlation between the elements of $\mathbf{Z}_i$, unlike sampling from $p(\mathbf{Z}|\mathbf{x}, \hat{\tau})$. This method also contains some previously considered choices of $h(\cdot)$ as special cases. If each observation is its own group and $\mu_{ij} = p(z_{ij} = 1|\hat{\tau}, \mathbf{x})$, then we have sampling from the conditional posterior for $\mathbf{Z}$, $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$. If the prior on π is a Dirichlet distribution (its natural conjugate prior), then $p(\mathbf{Z})$ has a Dirichlet-multinomial distribution. Hence, the above algorithm results in sampling from $p(\mathbf{Z})$ when $R = 1$ and $p(\mathbf{Z})$ is used in step 2.

Another version we have examined that allows for adaptive estimation of the importance sampling distribution is to set μ_j equal to the mean of the $p(\mathbf{Z}|\mathbf{x}_i, \hat{\tau})$ over $\mathbf{x}_i$'s in the j^{th} group, and take $\omega_j = cv_j/(\hat{p}_j(1 - \hat{p}_j) - v_j)$ where v_j is the variance of the $p(\mathbf{Z}|\mathbf{x}_i, \hat{\tau})$'s in the j^{th} group, $\hat{p}_j$ is their mean and c is an estimated tuning parameter. The *ad hoc* justification for this is that $\omega = v/(\pi(1 - \pi) - v)$ for random p's from a Beta(a, b) distribution where $\pi = a/(a + b)$ is its mean, v is its variance and $\omega = (a + b)^{-1}$. We chose c adaptively using (11). For reference, we call this sampling version Z–distance 2, or ZD–2, sampling. See Appendix C for another suggested variant of Z–Distance sampling.

The choice of the Dirichlet-multinomial distribution in step 2 is one of practicality and could be modified. This formulation of Z–distance sampling is a very general one that allows for a wide range of sampling schemes that cannot all be evaluated in a single study. Nevertheless, we put it forth as a general proposal given its intuitive appeal and unifying nature.

A simple version of the Z–distance method that we have evaluated in several simulations is the "Uniform Distance" method, or UD method, defined as follows. Set $R = G$, use the initial group assignments given by (12), and take $\mu_j = (1/G, \ldots, 1/G)\ \forall\ j$ and $\omega_j = 1/G\ \forall\ j$. The UD method has the following properties :
(1) It is adaptive in the sense that it uses information from $\hat{\tau}$ to form the initial grouping, but it does not require estimation of additional importance sampling distribution parameters. A modified form of the UD method, UD$_{mod}$, allows $R \geq G$ initial groups to be chosen by applying a non–parsimonious clustering algorithm. This appears to be useful when the data give only vague information about the true groupings.
(2) While $h(\hat{\mathbf{Z}}_M) \neq p(\hat{\mathbf{Z}}_M|\mathbf{x})$ as in the methods above, K can be chosen to control the probability of including $\hat{\mathbf{Z}}_M$ in the sample Q times for chosen Q ($Q = 5$, for example).
(3) The sampling is symmetric with respect to label-switching. That is, if $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are identical except for different labeling of the groups, then $h(\mathbf{Z}_1) = h(\mathbf{Z}_2)$ under the UD method of sampling. Property (3) provides an easy way of dealing with multimodality, discussed in more detail below.

3 Label-Switching and Multimodality

The "label–switching problem" refers to the fact that when the mixture components all have the same parametric form, the likelihood has the same value for different labelings of the groups:

$$\begin{aligned} &f\left(\mathbf{x}_i \mid \mathbf{z}_i = (z_{i1}, \ldots, z_{iG}), \theta = (\theta_1, \ldots, \theta_G), \pi = (\pi_1, \ldots, \pi_G)\right) \\ &= f\left(\mathbf{x}_i \mid \mathbf{z}_i = (z_{ip_1}, \ldots, z_{ip_G}), \theta = (\theta_{p_1}, \ldots, \theta_{p_G}), \pi = (\pi_{p_1}, \ldots, \pi_{p_G})\right), \end{aligned} \qquad (15)$$

where $(p_1, \ldots, p_G)$ is any permutation of $(1, \ldots, G)$. In general, there are up to $G!$ distinct labelings that result in the same value of $f(\mathbf{x}|z, \theta)$, and we say that these labelings are equivalent. There will be fewer than $G!$ distinct labelings when not all the parameter values are distinct. Note, in particular, that this is the case when two or more of the π_j's are zero. If $p(\theta, \pi)$ is symmetric with respect to label–switching, that is if $p\left((\theta_1, \ldots, \theta_G), (\pi_1, \ldots, \pi_G)\right) = p\left((\theta_{p_1}, \ldots, \theta_{p_G}), (\pi_{p_1}, \ldots, \pi_{p_G})\right)$, then $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ will have G! modes with equal mass. Because of this fact, we use a symmetric prior. This results in no loss of generality, because any non-symmetric prior can be reformulated into an equivalent symmetric prior with the same total mass on each set of equivalent points, so that I remains unchanged.

The defensive mixture method as described above produces a unimodal sampling distribution that can be a poor approximation to $p(\mathbf{Z}|\mathbf{x})$ under multimodality. The obvious modification is

$$h(\mathbf{Z}) = (1-\delta)\frac{1}{G!}\sum_b p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau}_b) + \delta p(\mathbf{Z}), \tag{16}$$

where b indexes the $G!$ permutations of the components of $\hat{\tau}$. The sampling is best done in a stratified manner, since this is both simpler and more efficient (Oh and Berger, 1993; Hesterberg, 1995): if K samples are planned, then $\mathbf{Z}$ is drawn from $p(\mathbf{Z})$ for $K\delta$ of the samples, and from each of the $p(\mathbf{Z}|\tau = \hat{\tau}_b, \mathbf{x})$'s for $K(1-\delta)/G!$ of the samples. We may still calculate δ_{opt} using (11).

As noted above, the UD method is already symmetric with respect to label–switching, so no modification is needed.

As an alternative to sampling at each mode, we have used the following method, which is applicable when the symmetric prior is employed. Assign each $\mathbf{Z}_k$ to an equivalence class containing all the $\mathbf{Z}$'s corresponding to an equivalent labeling in (15). Hence, the equivalence class to which $\mathbf{Z}_k$ belongs consists of $\mathbf{Z}_k$ plus all distinct permutations of the G columns of $\mathbf{Z}_k$. Denote the resulting equivalence classes by E_j, $j = 1, \ldots, N_E$, where N_E, the number of such equivalence classes, need not be known. Sampling from $h(z)$ can then be viewed as sampling from the E_j, and it follows that

$$\begin{aligned} I &= \sum_{j=1}^{N_E} \sum_{\mathbf{Z}_k \in E_j} L(\mathbf{x}|\mathbf{Z}_k) p(\mathbf{Z}_k) && (17) \\ &= \sum_{j=1}^{N_E} \#(E_j) L(\mathbf{x}|\mathbf{Z}_j) p(\mathbf{Z}_j), && (18) \end{aligned}$$

where $\mathbf{Z}_j$ is any $\mathbf{Z} \in E_j$ and $\#(E_j)$ is the number of $\mathbf{Z}$'s in E_j. It is easy to see that $\#(E_j) = G!/E0_j!$, where $E0_j$ is the number of groups with no members in it (i.e the number of columns consisting entirely of zeros) for $\mathbf{Z}$'s in the equivalence class E_j. Let $j(\mathbf{Z})$ denote the subscript of the equivalence class to which $\mathbf{Z}$ belongs. The importance sampling estimate is

$$\hat{I} = \sum_{k=1}^{K} \frac{G!}{E0_{j(\mathbf{Z}_k)}!} L(X|\mathbf{Z}_k) \frac{p(\mathbf{Z}_k)}{\sum_{\mathbf{Z} \in E_{j(\mathbf{Z}_k)}} h(\mathbf{Z})}. \tag{19}$$

In practice, the equivalence classes need not be enumerated. At each iteration, we sample $\mathbf{Z}_k$ from $h(\cdot)$ and then calculate $E0_{j(\mathbf{Z}_k)}$ and $\sum_{\mathbf{Z} \in E_{j(\mathbf{Z}_k)}} h(\mathbf{Z})$ for that iteration. The last step can take some computing effort. However, if $h(\cdot)$ is highly concentrated at one of the symmetry–based modes, $\sum_{\mathbf{Z} \in E_{j(\mathbf{Z}_k)}} h(\mathbf{Z})$ can be approximated by $h(\mathbf{Z}_k)$. This will be the case when the likelihood surface is peaked and δ is not close to 1.

Another method for handling the label–switching problem under a symmetric prior is the following. Instead of estimating I, estimate $\tilde{I}$ where

$$\tilde{I} = \int_S \frac{G!}{E0_{j(\mathbf{Z}_k)}!} L(\mathbf{x}|\mathbf{Z}) p(\mathbf{Z}) d\mathbf{Z} = \int \frac{G!}{E0_{j(\mathbf{Z}_k)}!} L(\mathbf{x}|\mathbf{Z}) p(\mathbf{Z}) 1[\mathbf{Z} \in S] d\mathbf{Z}, \tag{20}$$

where S represents any one of $G!$ equivalent "regions" of integration. "Regions" is put in quotes because some $\mathbf{Z}_k$'s will have fractional membership in more than one region. For example, when $G = 2$ and n is odd, $S = \{\mathbf{Z} : \sum_{i=1}^{(n-1)/2} z_{i2} \le (n-1)/2\}$. When n is even, $S = \{\mathbf{Z}_k : \sum_{j=1}^{n/2-1} z_{i2} \le n/2 - 1\}$, while the points $\mathbf{Z}$ with $\sum_{j=1}^{n/2} z_{k2} = n/2$ has a membership weight of .5 in S. Hence, we can sample near one mode using h_δ as constucted in Section 2, and any $\mathbf{Z}_k$ that falls outside of S for that mode contributes $I_k = 0$ to (7). When n is even and $\sum_j^{n/2} z_{k2} = n/2$, I_k is weighted by .5 in (7). When sampling is done mostly near the mode, few $\mathbf{Z}_k$'s fall outside S and little computing efficiency is lost in exchange for substantial simplicity.

With this particular choice of S, the proposed method seems likely to work best when the estimated proportions in the different groups are not too similar. The problem is similar to the label-switching problem in MCMC estimation of mixture models, and methods for dealing with this could also be adapted to the present setting. The most promising such methods essentially amount to different choices of S and algorithms for finding it, and the approaches of Celeux, Hurn and Robert (2000) and Stephens (2000) might be useful here also.

4 Simulations and Applications

Simulation studies were carried out to assess the performance of the DMIS and UD methods. Variations of the Z-distance method were applied in selected cases. Our simulations are divided into two groups. The first group consists of three small scale problems (Data Sets 1, 2, and 3) where performance can be assessed relative to the exact answer. The goals of this group of simulations were to assess the absolute and relative performance of these methods, to assess the sensitivity of the DMIS method to choices of δ, to assess the sensitivity of the

Table 1: Three Modes for the Likelihood Surface for Data Set 3.

-log lik	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\pi}$	$L(\mathbf{x}, \hat{\mathbf{Z}}_M) \times 10^{80}$
175.74	.169	.276	.155	2.48
175.78	.220	.151	.5	0.0013
175.91	.185	.532	0	11.3

UD method to the initial grouping, and to determine how the performance of each method varied depending on whether the likelihood was sharply peaked or flat.

In the second group of simulations, the methods are applied to three data sets (Data Sets 4, 5, and 6) with larger sample sizes (n=204). The goals of the second group of simulations were to assess the practical applicability of each method for larger samples, to determine whether there were any obvious breakdowns in the methods, to obtain a rough assessment of the accuracy of the methods for larger-sample problems, and to determine whether performance varied depending on the extent to which the likelihood was flat or peaked.

For all the simulation studies, the data were in the form of $\mathbf{X}_i$ successes in n_i trials where n_i was known. The fitted model was a two–component binomial mixture with uniform priors on the two mean parameters and on the mixing proportion parameter. We used $K = 1000$ for Data Sets 1, 2, and 3, and $K = 10,000$ for Data Sets 4, 5, and 6. This model was chosen for the simulation study in order to study the performance of the importance sampling methods for a particular application in molecular biology (Newton *et al*, 1998; Desai, 2000). Data Sets 1 and 2 are each a subset from one of two allelotype data sets (Barrett *et al*, 1996; and Shibagaki *et al*, 1994, respectively) where it is of interest to determine whether there exist two binomial components (Newton *et al*, 1998). Data Set 3 is simulated data from a single binomial component with mean 0.22 and n_i's that are the same as the n_i's in Data Set 2. [Should we show the data sets in an appendix? - AR, 5/12]

The resulting likelihood surfaces for τ for the three data sets range from markedly peaked to flat. The likelihood surface for τ for Data Set 3 is relatively flat with three modes of nearly equal height which correspond to disparate values of τ (Table 1). In Table 1, $\hat{\mathbf{Z}}_M$ is determined by (12) using the $\hat{\tau}$ indicated in the row. Note that the MLE is given in row 1, but the $\hat{\mathbf{Z}}_M$ derived from the MLE accounts for only 2.5% of the mass of I, while the $\hat{\mathbf{Z}}_M$ in the third row accounts for 11% of the mass. This shows that the MLE does not always translate to the best estimate of $\mathbf{Z}_M$.

Table 2 shows results for the defensive mixture method as a function of δ. We calculated δ_{opt} from (10). In each data set, δ_{opt} gave the minimum or near minimum variance for I, indicating that the theoretical arguments underlying its use translate to good performance in practice. As expected from its derivation, δ_{opt} performed best in Data Set 1 where the likelihood is very peaked and the assumption of only two distinct values for $L(\mathbf{x}|\mathbf{Z})$ is nearly satisfied. However, δ_{opt} still performed well in Data Set 3 where the assumption is not even approximately true. Mean–squared errors as a function of δ followed the same pattern. On the other hand, the empirical standard deviation of the I_k's was a better estimate of the (estimated) true standard deviation at δ_{opt} in Data Sets 2 & 3 than in Data Set 1. The results suggest that a δ in $[\delta_{opt}, 0.99]$ may provide the best balance of efficiency and estimated standard deviations that are accurate estimates of the true standard deviation.

Comparative results for the DMIS and UD methods are shown in Table 3, along with results for the UD method as a function of the initial grouping. For comparison, results from Monte Carlo estimation by sampling from the prior $(h(\tau) = p(\tau))$ are also shown. The variants UD-1, UD-2 and UD-3 in Table 3 refer to using the $\hat{\mathbf{Z}}_M$'s calculated from (12) using the three different $\hat{\tau}$'s in rows 1–3, respectively, in Table 1 as the initial groupings for the UD method. For comparison, the table entry $h(\tau) = p(\tau)$ refers to the result obtained by sampling from $p(\tau)$ for integration of (1). To assess the results, consider the scale of evaluation of evidence of Bayes factors in Jeffreys (1961) and Kass and Raftery (1995), according to which a Bayes factor between 1/3 and 3 constitutes evidence "worth no more than a bare mention." This suggests that we would want Bayes factors (ratios of two integrated likelihoods) to be off by no more than a factor of 3 with high probability. Some simple calculations suggest that this will be satisfied if the ratio of $\sqrt{\text{MSE}}$ to I is no more than about 0.2.

Overall, the results for the DMIS and UD methods were satisfactory. The bias was small, the coefficient of variation was well within the desired range for both methods and all data three sets, and the estimated standard errors were close to the empirical standard deviations, on average. In contrast, Monte Carlo estimation by sampling from the prior for the parameters τ was substantially less accurate on average; more samples would be needed for it to be minimally acceptable.

In Data Set 1 the DMIS method, with its ability to mimic the peaked posterior, had the best efficiency. The UD method and sampling from the prior for τ were less efficient but they did yield reasonably accurate standard errors. The DMIS and UD methods performed comparably in Data Set 2 for the modestly peaked likelihood, and the UD method was slightly better than the DMIS method for the flat likelihood in Data Set 3.

Table 2: Small Sample Simulation Results (N=17): Performance of defensive mixture importance sampling as a function of δ

Data set 1: markedly peaked likelihood
$I = 56.38 \times 10^{-76}$

δ	$\hat{I}$	$\bar{\text{SEM}}$	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
0	54.62	2.71	4.68	5.00
0.01	55.38	3.15	5.01	5.12
0.027	56.04	3.78	7.47	7.48
0.1	55.79	3.48	6.19	6.22
0.208*	55.21	3.16	4.44	4.60
0.5	57.49	5.13	6.43	6.52
0.9	56.61	7.52	6.84	6.84
0.99	59.57	18.84	11.36	11.79
1	66.55	52.23	78.89	79.54

Data set 2: peaked likelihood
$I = 51.29 \times 10^{-85}$

δ	$\hat{I}$	$\bar{\text{SEM}}$	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
0	44.27	3.78	5.38	8.85
0.01	54.13	11.42	63.76	63.83
0.1	50.27	5.93	13.53	13.57
0.5	49.29	4.62	5.73	6.07
0.586*	50.51	5.51	6.40	6.44
0.9	50.43	7.59	8.23	8.27
0.99	53.5	16.04	15.07	15.23
1	56.89	39.41	86.41	86.59

Data set 3: flat, multimodal likelihood
$I = 98.87 \times 10^{-80}$

δ	$\hat{I}$	$\bar{\text{SEM}}$	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
0	89.78	18.3	47.85	48.71
0.01	95.12	17.28	25.58	25.85
0.1	98.7	11.88	13.90	13.90
0.5	97.86	7.20	7.83	7.89
0.9	98.76	7.24	8.50	8.50
0.99	98.25	8.63	9.14	9.16
1*	97.62	9.53	9.54	9.62

NOTE: Each row represents results of 100 trials of the procedure with $K = 1000$ for each trial. $\bar{\hat{I}}$ is the mean of $\hat{I}$ over the 100 trials, $\bar{\text{SEM}}$ is the mean of the standard errors of $\hat{I}$ over the 100 trials, and SD($\hat{I}$) is the empirical standard deviation of the 100 $\hat{I}$'s (taken to be an estimate of the true standard deviation of $\hat{I}$ for comparisons with $\bar{\text{SEM}}$).

* This is δ_{opt} from (10).

Table 3: Small Sample Simulation Results (N=17): Comparison of Methods

Data set 1: markedly peaked likelihood $I = 56.38 \times 10^{-76}$				
Method	$\hat{I}$	SĒM	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
DMIS	55.21	3.16	4.44	4.60
UD	57.10	7.38	8.50	8.53
$h(\tau) = p(\tau)$	56.78	13.12	13.34	13.35
Data set 2: peaked likelihood $I = 51.29 \times 10^{-85}$				
Method	$\hat{I}$	SĒM	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
DMIS	50.51	5.51	6.40	6.44
UD	52.16	6.13	6.18	6.24
$h(\tau) = p(\tau)$	52.17	9.77	10.44	10.48
Data set 3: flat, multimodal likelihood $I = 98.87 \times 10^{-80}$				
Method	$\hat{I}$	SĒM	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
DMIS	97.62	9.53	9.54	9.62
UD-1	94.98	6.20	6.52	7.67
UD-2	98.13	6.64	7.12	7.16
UD-3	98.27	6.63	7.17	7.20
$h(\tau) = p(\tau)$	98.66	15.37	13.70	13.70

NOTE: Each row represents results of 100 trials of the procedure with $K = 1000$ for each trial. The columns are defined as in Table 2.

UD-b refers to the Uniform Distance method taking the initial grouping to be $\hat{Z}_M$ calculated from (12) using $\hat{\tau}$ from row b of Table 1.

The three larger sample simulation data sets each consisted of 204 proportions. Data Set 4 consists of 12 replicates of Data Set 1, Data Set 5 consists of 12 replicates of Data Set 3, and Data Set 6 consisted of 204 replicates of the proportion 8/40. Data Sets 4 and 5 were constructed in this way in order to be able to get bounds on I using the exact values of I for Data Sets 1 and 3, and in order to assess the impact of increasing the sample size while keeping the shape of the likelihood surface fixed. To obtain bias estimates for Data Set 5, the true I was taken to be the value of $\hat{I}$ resulting from importance sampling integration of (3) by drawing 5 million samples directly from the prior on the parameters, $p(\theta, \pi)$. This appears to provide a reliable answer for the flat likelihood in this example. Data Set 6 is an artificial example with an easily calculated exact answer that should be "easy" for any viable candidate method. Data Set 6 serves as a gauge for assessing whether a method meets a minimum performance standard.

Results for the larger sample sizes are shown in Table 4. The results show that the 2-component DMIS method performs very well for the markedly peaked likelihoods in Data Sets 4 and 6, but performs poorly for the flat likelihood (Data Set 5) compared to sampling from $p(\tau)$. Because the likelihood for Data Set 5 has three modes of comparable height, a 4-component DMIS sampling method was tested here, with components corresponding to each mode and to the prior as in (13). Taking each of the four δ's (component sampling weights) in this sampling scheme to be 0.25 results in a great improvement in the bias as well as improvement in the variance of $\hat{I}$ compared to the 2-component DMIS importance sampling distribution (Table 4, first and second rows under Data Set 5). When we attempt to solve (14) in order to minimize the variance resulting from the 4–component importance sampling distribution, we find no solution in $[0,1]^{\otimes 4}$. Without this constraint, the solution puts δ_3 at 0.012, $\delta < 0$, $\delta_2 > 1$ and $\delta_3 > 1$. Hence, we set $\delta_3 = 0.012, \delta = 0$ and split the remaining mass between δ_2 and δ_3. This was very effective in reducing $\text{var}(\hat{I})$, producing a 54–fold decrease in $\text{var}(\hat{I})$ relative to the equal weight case (Table 4, third row under Data Set 5). However, the estimate of I was biased downward, and only 50% of the mass was "found". Still, the MSE was significantly improved. We also adjusted the δ_i's to allow $\delta = 0.05$ (versus $\delta = 0$ in the previous simulation). The result was an increase in the variance and decrease in the bias, giving an overall decrease of 5% in the MSE (Table 4).

The naive UD method with $R = G$, as described in Section 2.2, does not perform well compared to the DMIS method for the markedly peaked likelihood, nor does it perform as well as sampling from $p(\tau)$ for the flat likelihood in Data Set 5. (Note that the UD method and the DMIS method with $\delta = \delta_{opt}$ coincide in this case.) However, the ZD–2 method

Table 4: Larger Sample Simulation Results (N=204)

Data set 4: Peaked Likelihood, $I \in [37.4, 486.1] \times 10^{-870}$				
Method	$\hat{I}$	S$\bar{\text{E}}$M	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
DMIS, $\delta = 0.078$	65.32	0.31	0.30	—
UD	66.62	26.01	30.91	—
ZD-2[1]	65.32	0.11	0.13	—
$h(\tau) = p(\tau)$	78.82	37.13	39.10	—
Data set 5: Flat, Multimodal Likelihood, $I \in [0.25, 1762] \times 10^{-920}$				
Method	$\hat{I}$	S$\bar{\text{E}}$M	SD($\hat{I}$)	$\sqrt{\text{MSE}}^*$
2–component DMIS $\delta = .5$	12.56	7.65	30.44	31.16
4–component DMIS[2]	19.98	8.13	24.41	24.42
4–component DMIS$^3_{opt}$	9.35	2.08	3.22	10.41
4–component DMIS, $\delta = .05$[4]	14.59	3.63	6.54	8.03
2–component DMIS, $\delta = \delta_{opt} = 1$; UD[5]	11.83	6.02	14.04	15.87
mixture of 3 UDs[6]	15.37	4.31	6.59	7.64
UD$_{mod}$, R=3[7]	17.11	3.36	4.42	4.09
$h(\tau) = p(\tau)$	18.85	2.88	2.63	2.63
Data set 6: One Group; Peaked Likelihood, $I = 20.95 \times 10^{-1778}$				
Method	$\hat{I}$	S$\bar{\text{E}}$M	SD($\hat{I}$)	$\sqrt{\text{MSE}}$
DMIS, $\delta_{opt} = 0.7716$	20.99	0.24	0.24	0.24
UD	21.06	0.49	0.40	0.41
$h(\tau) = p(\tau)$	21.52	6.58	5.89	5.92

NOTES: Each row represents 50 trials with $K = 10000$ for each trial.

* estimated based on assuming I = 19.240 obtained as $\hat{I}$ from 5 million samples from $p(\tau)$. The standard error of this estimate is 0.001

[1] ZD-2 sampling as described in Section 2.2.

[2] The IS distribution is $h(\mathbf{Z}) = \delta_1 p(\mathbf{Z}|x, \tau = \hat{\tau}_1 + \delta_2 p(\mathbf{Z}|x, \tau = \hat{\tau}_2 + \delta_3 p(\mathbf{Z}|x, \tau = \hat{\tau}_3 + \delta p(\mathbf{Z})$, as in (13), where the $\hat{\tau}_i$'s correspond to the three modes of the likelihood surface (Table 1). $\delta = \delta_i = 0.25, i = 1, 2, 3$.

[3] h is the same as above, except that δ and the δ_i's were chosen adaptively. The solution to (14) puts $\delta_3 = 0.012,\ \delta_2 > 1, \delta_1 > 1$ and $\delta < 0$, so $\{\delta, \delta_1, \delta_2, \delta_3\} = \{0, .494, .494, 0.012\}$ was used.

[4] Like the case directly above, except that δ was set to 0.05: $\{\delta, \delta_1, \delta_2, \delta_3\} = \{.05, .47, .47, .012\}$.

[5] In this case the 2–component DMIS with $\delta = \delta_{opt} = 1$ is equivalent to the UD method, since the latter puts all observations into one group initially ($R = 1$).

[6] Three versions of the UD method were mixed in equal proportions. Each of the UD initial groupings with R=2 was formed from one of the $\hat{Z_M}$'s calculated from (12) using the three $\hat{\tau}$'s in (Table 1).

[7] Three initial groups for the UD method (R=3) were created using an *ad hoc* clustering method (inspecting the values of $p(\mathbf{z}_i|\mathbf{x}_i, \tau = \hat{\tau}_{MLE})$ and dividing them into three groups according to their proximity.)

and the simple UD$_{mod}$ method brought large gains in efficiency in each case, respectively. Specifically, using three initial groups for the UD$_{mod}$ method (R=3) reduced the estimated MSE by a factor of 10 for Data Set 5, while use of ZD-2 sampling, as described in Section 2.2, resulted in a variance reduction for $\hat{I}$ by a factor greater than 5×10^4 (Table 4). For the contrived Data Set 6, both the naive UD method and the DMIS method performed well and considerably outperformed sampling from the prior. Sampling from the prior peformed best for Data Set 5, emphasizing that one should always investigate simple methods when the integrand has simplifying features. Most applications will not present such simple likelihoods, however, and adaptive methods such as those proposed here may be needed.

5 Discussion

We have proposed a general approach to approximating integrated likelihoods for finite mixture models. The basic idea is to make the unobserved group memberships the argument of the integral, and then to use importance sampling in terms of the group memberships to evaluate the integral. The tricky part of this is finding a good importance sampling function, especially since the dimension of the resulting integral is high in most cases. We propose using a mixture distribution a second time in the problem, this time as importance sampling function, and we outline two ways of doing this: defensive mixture importance sampling, and the Z–distance method. We propose choosing the mixture parameters of the importance sampling function adaptively so as to minimize the variance of the approximated integrand, and we develop a simple way of making this choice.

The resulting method is easy to implement, essentially involving only simple multinomial sampling. It does not become much more complex as the complexity of the mixture model increases, and extends easily to more complicated mixture models. It is based on independent simulation, and so does not encounter problems of convergence due to dependence between consecutive samples. It also produces a standard error and so to some extent is self-monitoring, although care must be taken when using this. We also propose a way of dealing with the label-switching problem that can plague inference for mixture models. The method can also allow one to make use of analytic approximations that are valid when the group memberships are known (such as the Laplace method and BIC), even though these are not valid for the mixture model itself.

The method seems to have some advantages over the main alternative methods that have been proposed. Analytic approximations such as Laplace and BIC are not valid for finite mixture models. The number of parameters in mixture models tends to outrun the capabilities of numerical quadrature rather quickly. And our methods seem easier to implement than MCMC, also addressing the convergence and label-switching problems that can affect MCMC for mixture models. For further discussion of the relative merits of importance sampling and MCMC in a different context, see Stephens and Donnelly (2000).

The basic idea of our method is not limited to mixture models, and can be applied more generally to models that can be expressed simply in terms of latent or missing data, i.e. models that lend themselves to the EM algorithm or to data augmentation. These include simple hierarchical models, variance component models, and multiple imputation models for missing data.

Various other approaches have been proposed for approximating integrated likelihoods for models of this kind using the output of the EM algorithm or similar results. Cheeseman and Stutz (1995) proposed the estimator

$$\hat{I}_{CS} = L(\mathbf{x}|\hat{\mathbf{Z}})p(\hat{\mathbf{Z}})/h(\hat{\mathbf{Z}}), \tag{21}$$

where $h(\hat{\mathbf{Z}}) = p(\hat{\mathbf{Z}}|\mathbf{x},\hat{\tau})|_{\mathbf{Z}=\hat{\mathbf{Z}}}$. Our discussion here sheds some light on this estimator of the integrated likelihood. Recall that $I = \int L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})d\mathbf{Z}$, and that the importance sampling estimator of I based on the importance sampling function $h(\cdot)$ is $\hat{I} = K^{-1}\sum_{k=1}^{K} L(\mathbf{x}|\mathbf{Z}_k)p(\mathbf{Z}_k)/h(\mathbf{Z}_k)$. Now the integrated likelihood can be rewritten as

$$I = \int v(\mathbf{Z})h(\mathbf{Z})d\mathbf{Z}, \tag{22}$$

where $v(\mathbf{Z}) = L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})/h(\mathbf{Z})$. The importance sampling estimator is then seen to be the simple Monte-Carlo integration estimator

$$\hat{I} = K^{-1}\sum_{k=1}^{K} v(\mathbf{Z}_k). \tag{23}$$

The Cheeseman-Stutz estimator is then just

$$\hat{I}_{CS} = v(\hat{\mathbf{Z}}). \tag{24}$$

From (22), (23) and (24), it is apparent that instead of the simulation-consistent approach of integrating over values of $\mathbf{Z}$ simulated from $h(\cdot)$, the Cheeseman-Stutz estimator proceeds by replacing $\mathbf{Z}$ by $\hat{\mathbf{Z}}$ in the integrand. This is not valid in the present context, and the Cheeseman-Stutz estimator will be biased in general, even with an infinite amount of simulation. Note that the published derivation of the Cheeseman-Stutz estimator was based on a Laplace approximation that is not valid in general for mixture models.

Chickering and Heckerman (1996) proposed a modification to the Cheeseman-Stutz estimator to take account of the possibility that the dimension of the parameter space for the completed data set $(\mathbf{x}, \mathbf{Z})$ may be different from that for the imcomplete data. The need for such an adjustment arises only because of the Laplace method used to derive the Cheeseman-Stutz estimator which, as we have seen, is in any event invalid for general mixture models. No such adjustment is needed in our importance sampling approach.

Biernacki, Celeux and Govaert (2000) have proposed the Integrated Completed Likelihood (ICL), which is similar to the Cheeseman-Stutz approach in that it is based on a single value of $\mathbf{Z}$, but replaces the elements of $\mathbf{Z}$ by their most likely values given the data and $\hat{\tau}$ (rather than by their expected values as in Cheeseman-Stutz), and then integrates the resulting completed likelihood over τ. Thus, Biernacki *et al.* (2000) report

$$p(\mathbf{x}|\hat{\mathbf{Z}}_M) = \int p(\mathbf{x}|\hat{\mathbf{Z}}_M, \tau) p(\tau|\hat{\mathbf{Z}}_M) d\tau,$$

where $\hat{\mathbf{Z}}_M$ is given by (12). This is clearly not valid as an approximation to the integrated likelihood, but Biernacki *et al.* (2000) do not claim that it is, and argue for it instead in its own right as the solution to a scientific problem that is of interest in some contexts.

Wei and Tanner (1990) proposed importance sampling using $h(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$, a method they called Poor Man's Data Augmentation–2 (PMDA–2). They used this for parameter estimation, and not, as here, for integrated likelihoods. However, it could be used for integrated likelihoods, and then it would be a special case of DMIS, with $\delta = 0$. As an importance sampling function, this may miss important regions away from the MLE, and the resulting estimator of the integrated likelihood may have high variance (or even, if $\mathbf{Z}$ is continuous, infinite variance). In Wei and Tanner's (1990) proposed PMDA–1 method, the importance sampling weights are all set to be equal. This greatly reduces the variance, but the resulting estimator is biased, even asymptotically (i.e. with an amount of simulation that tends to infinity).

The methods we propose here are in the spirit of the nonparametric importance sampling method of Zhang (1996), in the sense that we obtain an adaptive estimate of the most efficient importance sampling distribution. Adaptive procedures for estimating an intractable $p(\mathbf{Z})$ are described by Evans and Swartz (1995), Oh and Berger (1993) and Givens and Raftery (1996). Our goal here is to go beyond estimating $p(\mathbf{Z})$ and estimate an importance sampling distribution which is closer to the ideal distribution $p(\mathbf{Z}|\mathbf{x})$ than to $p(\mathbf{Z})$. In addition, we have addressed the issue of accurate estimation of the standard error of $\hat{I}$.

Raghavan and Cox (1998) proposed a different adaptive algorithm for estimating the mixing parameter δ in defensive mixture importance sampling. Their method is designed for importance sampling problems where there is more than one estimand of interest. They employ a complex minimization and reweighting scheme that allows the researcher to try to minimize the asymptotic variances of several importance sampling estimators based on the same randomly sampled values. We can avoid using computational minimization techniques by using the much simpler optimal δ result in Theorem 2, because we are interested only in calculating the integrated likelihood.

We have explored various possible versions of DMIS and UD, finding that in situations where the likelihood is not highly peaked, better performance can be achieved by adding components to the defensive mixture, or by using mixtures of UD methods. It seems plausible that the method could be improved by developing a systematic way of determining when additional components are needed in the defensive mixture, and adding them automatically. One possible approach is adaptive. Consider, for example, $h(\mathbf{Z})$ to be a mixture of functions of some form such as (13) or, more parsimoniously, location- and/or scale-shifted versions of $p(\mathbf{Z}|\hat{\mathbf{Z}})$. In a first stage, two-component DMIS would be used to generate a first (weighted) sample from $p(\mathbf{Z}|\mathbf{x})$. Then, using the EM algorithm, a mixture of functions of $\mathbf{Z}$ of the selected form could be fit to the resulting sample, for example using some criterion such as BIC to choose the number of components and the selected functional form. In this way additional components would be added automatically to the defensive mixture if needed to fill out areas that are underrepresented in the first iteration. This adaptive process could be continued iteratively, although it seems likely that the improvement would be small past a few iterations.

Owen and Zhou (2000) discussed the use of control variates to improve the performance of defensive mixture sampling for integration. The control variate method provided impressive gains in efficiency in the examples therein. However, our trials of the control variates did not improve the performance of the DMIS method for integrated mixture likelihoods when the fitted G was greater than the true G (results not shown). More research is needed to determine how the techniques proposed by Owen and Zhou may complement those described here.

A more direct method for sampling $\mathbf{Z}$'s that are moderately close to Z_M is embodied in the Z–distance sampling approach introduced here. This method is a promising avenue for further research. The z_i's can be sampled in a designed, dependent manner that reflects obvious features of the data, and this may well have the potential to lead to improved methods. For example, the initial groups can be based upon the fact that particular pairs of points are much more likely to be in the same group than other pairs. Furthermore, in its more complex forms, Z-distance sampling allows tuning of its Dirichlet-multinomial sampling distributions to the data. The simplest form of Z-distance sampling, the Uniform Distance method with $R = G$, appears to be adequate for small samples and moderately peaked likelihoods. However, it uses little information from the data and may be too inefficient for many problems. For our large-sample flat likelihood test case (Data Set 5), was improved markedly by using more data information simply by choosing $R > G$ initial groups via a non–parsimonious initial clustering step. This method may be particularly useful when models with G greater than the true G will be fitted, resulting in flat likelihood surfaces with many minor modes. A problem for future research is to find a method for choosing the number of initial groups to optimize the MSE, say.

For peaked likelihoods, ZD-2 sampling performed as well as the DMIS method, using (10) to tune the importance sampling distribution parameters. The Z–distance sampling method proposed here allows very flexible sampling schemes. Even more flexible sampling schemes could be obtained by allowing the groupings to vary with k and/or allowing the assignments in the $j+1, j+2, ...R$ groups to be conditional on the assignments in the $1, \ldots, j$ groups.

One place where difficulties might arise with our variance–optimization method is in obtaining adequate initial estimates of I and Z_M. However, the need for some sort of information about the integrand is a general requirement for optimizing an importance sampling technique, not a shortcoming of our method in particular. In our examples, we had no trouble finding a suitable initial $\hat{I}$ by taking $\delta = .5$ in the 2–component DMIS method. Nevertheless, it is conceivable that the 2–component DMIS method might fail completely for some data/model combinations. A simple Laplace-based method such as BIC, even if inadequate as a final value, may be good enough to be used as an initial value.

The choice of a suitable importance sampling method (with a suitable number of components when the DMIS method is used) may be easier if the data analyst has some knowledge of the likelihood surface. For complex problems, this can be obtained, for example, by examining the surface using the EM algorithm or other mode–finding routine on a grid of starting points. This also provides a search for $\hat{\mathbf{Z}}_M$ by applying (12) at each located mode.

No single importance sampling variant can be expected to work for every data set, and some knowledge of the integrand is required in order to choose an appropriate variant. However, the small set of variants presented here seems to hold some promise of being useful over a reasonably wide range of integrands. The methods studied here are also simple, and we provide theoretical insights into constructing good importance sampling distributions through Theorems 1 and 2. Like Stephens and Donnelly (2000), we conclude that importance sampling deserves further development as a viable, potentially simpler alternative to MCMC methods for some problems.

APPENDIX

A. Proof of Theorem 1 For this proof, we redefine $\tau = (\pi_1, \ldots, \pi_{G-1}, \theta')'$ so that the information matrices can be assumed to be positive definite. Assume that the following conditions hold:

(1) τ is r–dimensional with prior density $p(\tau)$ that is continuous at τ_0, and $p(\tau_0) > 0$.
(2) $\hat{\tau}_n \xrightarrow{p} \tau_0$ where τ_0 is a point in $[0,1]^{G-1} \times \Theta$.
(3) There exists a function $\tilde{M}(\mathbf{X};\tau_0)$ with the property that given any $\epsilon > 0$

$$\left| \frac{\delta^2}{\delta\tau^2} \log f_{X,Z}(\mathbf{X}_i, \hat{z}(\mathbf{X}_i,\tau_0)|\tau) - \frac{\delta^2}{\delta\tau^2} \log f_{X,Z}(\mathbf{X}_i, \hat{z}(\mathbf{X}_i,\tau_0)|\tau_0) \right| < \tilde{M}(\mathbf{X}_i;\tau_0)$$

with $E[\tilde{M}(\mathbf{X}_i;\tau_0)] < \epsilon$ whenever $|\tau - \tau_0| < \delta$ for some δ.
(4) The functions $\hat{z}_{ij}(\mathbf{X}_i;\tau)$ are simultaneously continuous at τ_0 with probability going to 1;
(5) There exists a function $M(\mathbf{X}_i;\tau_0)$ with the property that given any $\epsilon > 0$

$$\left| \frac{\delta^2}{\delta\tau^2} \log f_X(\mathbf{X}_i|\tau) - \frac{\delta^2}{\delta\tau^2} \log f_X(\mathbf{X}_i|\tau_0) \right| < M(\mathbf{X}_i;\tau_0)$$

with $E[M(\mathbf{X}_i;\tau_0)] < \epsilon$ whenever $|\tau - \tau_0| < \delta$ for some δ.

We suppress some subscripts for ease of notation in the proof. Partition the ratio of interest as follows:

$$\begin{aligned} \frac{p_n(\hat{z}(\mathbf{X}^{(n)})|\mathbf{X}^{(n)})}{p_n(\hat{z}(\mathbf{X}^{(n)})|\mathbf{X}^{(n)}, \tau = \hat{\tau}_n)} &= \frac{\int p(\mathbf{X}^{(n)}, \hat{z}(\mathbf{X}^{(n)})|\tau)p(\tau)d\tau}{p(\mathbf{X}^{(n)}, \hat{z}(\mathbf{X}^{(n)})|\tau = \hat{\tau})} \times \frac{p(\mathbf{X}^{(n)}|\tau = \hat{\tau})}{\int p(\mathbf{X}^{(n)}|\tau)p(\tau)d\tau} \\ &\equiv \frac{N_1}{D_1} \times \frac{N_2}{D_2}. \end{aligned} \tag{25}$$

The proof is similar to that of Walker (1969) to show that

$$\frac{N_1}{n^{r/2} D_1 p(\hat{\tau})} \xrightarrow{p} (2\pi)^{r/2} |\tilde{J}(\tau_0)|^{1/2}. \tag{26}$$

Walker's θ and $f(\mathbf{X}_i|\theta)$ are replaced by τ and $f_{X,Z}(\mathbf{X}_i, \hat{z}_i(\mathbf{X}_i,\hat{\tau})|\tau)$, respectively, and a few extra steps are needed to show that $\hat{z}_i(\mathbf{X}_i,\hat{\tau})$ can be replaced by $\hat{z}_i(\mathbf{X}_i,\tau_0)$ in determining the limit. Briefly, the integrand in N_1 is expanded about $\hat{\tau}$ to obtain

$$\begin{aligned} N_1 &= \int \exp(\log(p(\mathbf{X}^{(n)}, \hat{z}(\mathbf{X}^{(n)})|\tau)p(\tau)))d\tau \qquad (27) \\ &= p(\mathbf{X}^{(n)}, \hat{z}(\mathbf{X}^{(n)})|\tau = \hat{\tau})p(\hat{\tau}) \int \exp\{\frac{1}{2}(\tau - \hat{\tau})'\tilde{C}_n(\hat{\tau};\hat{\tau})(\tau - \hat{\tau})[1 + R_{1n} + R_{2n}]\}d\tau, \end{aligned}$$

where $\tau^* \in (\hat{\tau}, \tau)$,

$$\begin{aligned}
\tilde{C}_n(\hat{\tau}; \tau^*) &= \frac{\partial^2}{\partial \tau^2} \log p(\mathbf{X}^{(n)}, \hat{z}(\mathbf{X}^{(n)})|\tau)|_{\tau^*}, \\
R_{1n} &= \frac{\log p(\tau) - \log p(\hat{\tau})}{\frac{1}{2}(\tau - \hat{\tau})' \tilde{C}_n(\hat{\tau}, \hat{\tau})(\tau - \hat{\tau})}, \text{ and} \\
R_{2n} &= \frac{(\tau - \hat{\tau})'[\tilde{C}_n(\hat{\tau}, \tau^*) - \tilde{C}_n(\hat{\tau}, \hat{\tau})](\tau - \hat{\tau})}{(\tau - \hat{\tau})' \tilde{C}_n(\hat{\tau}, \hat{\tau})(\tau - \hat{\tau})}.
\end{aligned}$$

Here we have used the fact that

$$\frac{\partial}{\partial \tau} p(\mathbf{X}^{(n)}, \hat{z}(\mathbf{X}^{(n)})|\tau) = 0 \text{ at } \hat{\tau}.$$

The assumptions ensure that R_{1n} and R_{2n} converge in probability to zero, uniformly over τ in a neighborhood of τ_0, and that the integrand is negligible outside this neighborhood. Also, $|\tilde{C}_n(\hat{\tau}, \hat{\tau})|/n^r \xrightarrow{p} |\tilde{J}(\tau_0)|$ by continuous mapping and (26) follows. Similarly, $D_2/(N_2 n^{n/2} p(\hat{\tau})) \xrightarrow{p} (2\pi)^{r/2} |J(\tau_0)|^{1/2}$, was shown by Walker (1969), which establishes the theorem.

B. Proof of Theorem 2

Let j index the G^n values of $\mathbf{Z}$, and let $h_{\delta j}$ denote the probability of $\mathbf{Z}_j$ under a generic importance sampling distribution $h(\cdot)$ that depends on δ in some manner. Also let k_j be the number of times $\mathbf{Z}_j$ is drawn in a sample of size K. Note that the variance of $\hat{I}$ is given by

$$\begin{aligned}
&\text{var}(\hat{I}) \equiv v_1(\delta) = \text{var}(\frac{1}{K} \sum_{j=1} \frac{L(\mathbf{x}|\mathbf{Z}_j) p(\mathbf{Z}_j) k_j}{h_{\delta j}}) \\
&= \frac{1}{K} (\sum_j \frac{L(\mathbf{x}|\mathbf{Z}_j)^2 p(\mathbf{Z}_j)^2 (1 - h_{\delta j})}{h_{\delta j}} - \sum_{j \neq i} L(\mathbf{x}|\mathbf{Z}_j) p(\mathbf{Z}_j) L(\mathbf{x}|\mathbf{Z}_i) p(\mathbf{Z}_i)),
\end{aligned} \qquad (28)$$

since $(k_1, \ldots, k_J)$ is multinomial $(K, (h_{\delta 1}, \ldots, h_{\delta G^n}))$. Note that omitting the dependence on δ and minimizing (28) with respect to each h_j results in

$$h_j = L(\mathbf{x}|\mathbf{Z}_j) p(\mathbf{Z}_j) / \sum_k L(\mathbf{x}|\mathbf{Z}_k) p(\mathbf{Z}_k) = p(\mathbf{Z}_j|\mathbf{x}),$$

the posterior probability of $\mathbf{Z}_j$. This shows that the unknown $p(\mathbf{Z}|\mathbf{x})$ is the optimal importance sampling distribution in terms of minimizing $\text{var}(\hat{I})$.

Returning to h's of known form depending on δ, a reasonable simplifying assumption at this point is to lump $\mathbf{Z}$'s with the same or nearly the same values of $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ into a single equivalence class and take $h(\mathbf{Z})$ to be the same for all $\mathbf{Z}$'s in an equivalence class. Because there is a class of points with negligible mass for $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$, we found it reasonable to approximate the minimization problem by assuming just two equivalence classes corresponding to the "non-negligible" and "negligible" classes of points, respectively. Suppose there are c_1 elements in the "non-negligible" class (class 1) and c_2 elements in the "negligible" class (class 2) , with f_1 and f_2 being the respective values of $L(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})$ and $h_{\delta 1}$ and $h_{\delta 2}$ being the respective values of h. Then,

$$v_1(\delta) = \frac{c_1 f_1^2(1-h_{\delta 1})}{h_{\delta 1}} + \frac{c_2 f_2^2(1-h_{\delta 2})}{h_{\delta 2}} + \text{constant}. \tag{29}$$

Now, minimizing (29) with respect to $h_{\delta 1}$, subject to the constraint $c_1 h_{\delta 1} + c_2 h_{\delta 2} = 1$ gives

$$h_{\delta 1} = \frac{f_1}{c_1 f_1 + c_2 f_2} = \frac{L(\mathbf{x}|\mathbf{Z}_M)p(\mathbf{Z}_M)}{I} = h(\mathbf{Z}_M). \tag{30}$$

The last equality follows from the fact that $\mathbf{Z}_M$ must be in the non-negligible class of points. Note that (30) does not depend on the actual form of $h(\cdot)$. When $h(\cdot)$ has the form of (8), (10) follows from (30).

One of the goals of this research is to find an importance sampling method whose precision can be accurately monitored using the sample variance of the I_k's. Hence, it is of interest to choose h to minimize or nearly minimize $v_2(\delta) \equiv \text{var}(\widehat{\text{var}}(\hat{I}))$. In this subsection, we show that $h_{\delta 1}$ in (30) also minimizes $v_2(\delta)$ under the conditions above. Specificallyy,

$$\begin{aligned} \text{var}(\widehat{\text{var}}(\hat{I})) &= \text{var}(\frac{1}{K}\sum_{k=1}^{K}(I_k - \hat{I})^2) & (31) \\ &= \frac{\mu_4 - \mu_2^2}{K} - \frac{2(\mu_4 - 2\mu_2^2)}{K^2} + \frac{\mu_4 - 3\mu_2^2}{K^3}, & (32) \end{aligned}$$

where μ_j represents the j^{th} central moment of the distribution producing the IID I_k's (Cramer, 1946). Let $y_k = 1$ if $\mathbf{Z}_k$ is in class 1 and let $y_k = 0$ if $\mathbf{Z}_k$ is in class 2. Then

$$\begin{aligned} I_k = L(\mathbf{x}|\mathbf{Z}_k)\frac{p(\mathbf{Z}_k)}{h(\mathbf{Z}_k)} &= \frac{f_1}{h_{\delta 1}} y_k + \frac{f_2}{h_{\delta 2}}(1 - y_k) & (33) \\ &= \left(\frac{f_1}{h_{\delta 1}} - \frac{f_2}{h_{\delta 2}}\right) y_k + \frac{f_2}{h_{\delta 2}} & (34) \\ &\equiv M_1(h_{\delta 1}) y_k + M_2(h_{\delta 1}). & (35) \end{aligned}$$

Hence,

$$\mu_2 = M_1^2 pq, \text{ and } \mu_4 = M_1^4 pq(1 - 3pq), \tag{36}$$

where $p = c_1 h_{\delta 1} = E[y_k]$ and $q \equiv 1 - p$. This gives

$$\text{var}(\widehat{\text{var}}(\hat{I})) \propto (K-1)M_1^4 pq - 2(2K-3)M_1^4 p^2 q^2. \tag{37}$$

It can be verified directly that the $h_{\delta 1}$ in (30) is a zero of the derivative of (37) and that it represents a minimum.

C. Z–distance Sampling Using a Dirichlet-Multinomial distribution

The Dirichlet-Multinomial distribution is a compound distribution that describes the marginal distribution of $\mathbf{X}$ when $\mathbf{X}|p$ is Multinomial$(n, p = (p_1, \cdots, p_G))$, and p has a Dirichlet distribution whose joint density is given by

$$C p_1^{\alpha_1 - 1} \ldots p_G^{\alpha_G - 1}, \tag{38}$$

where $p_i \geq 0$, $\sum p_i = 1$, $S \equiv \sum \alpha_i$ and $\alpha_i > 0$. Then the marginal mean of $\mathbf{X}$ is $n\mu$ where the mean parameter $\mu = (\alpha_1, \ldots, \alpha_G)/S$; and the dispersion parameter for $\mathbf{X}$ is $\omega = 1/S$. The probability mass function for $\mathbf{X}$ is

$$\text{Prob}[\mathbf{X} = (x_1, \ldots, x_G)] = \frac{n!\Gamma(\omega^{-1}) \prod_i \Gamma(\mathbf{x}_i + \mu_i \omega^{-1})}{\prod_i \mathbf{x}_i! \prod_i \Gamma(\mu_i \omega^{-1}) \Gamma(n + \omega^{-1})}. \tag{39}$$

To facilitate the Z-distance sampling described in section 2, it is useful to note that the number of observations falling into group j is beta-binomial with probability mass function

$$\begin{aligned} \text{Prob}[\mathbf{X}_j = c] &= \frac{n!\Gamma(\omega^{-1})\Gamma(c + \mu_1\omega^{-1})\Gamma(n - c + (1-\mu_1)\omega^{-1})}{c!(n-c)!\Gamma(\mu_1\omega^{-1})\Gamma((1-\mu_1)\omega^{-1})\Gamma(n + \omega^{-1})} & (40) \\ &= \frac{n!\Gamma(S)\Gamma(c + \alpha_1)\Gamma(n - c + S - a\alpha_1)}{c!(n-c)!\Gamma(\alpha_1)\Gamma(S - \alpha_1)\Gamma(n + S)}. & (41) \end{aligned}$$

Given $\mathbf{X}_1 = c$, the conditional distribution of $(\mathbf{X}_2, \ldots, \mathbf{X}_G | \mathbf{X}_1 = c)$ is Dirichlet-Multinomial with mean parameter $\mu' = (\mu_2, \ldots, \mu_G)/\sum_{j \neq 1} \mu_j$, and dispersion parameter $\omega' = \omega/\sum_{j \neq 1} \mu_j$. Hence, the distribution of the n observations into the G groups can be done as a sequence of beta-binomial samples. This is particularly easy when $\mu = (1/G, \ldots, 1/G)$ and $\omega = 1/G$ for the Uniform Distance method. The Uniform Distance method is a variant of our Z–distance method that incorporates a limited amount of information from $p(\mathbf{Z}|\mathbf{x}, \tau = \hat{\tau})$. When necessary, improved performance might be achieved by taking $\mu_j = (1-\alpha)(0, \ldots, 1, \ldots, 0)' + \alpha\hat{\mu}$, where the first component vector has a 1 in the j^{th} position, and the k^{th} element of the second component vector is given by

$$\hat{\mu}_{jk} = \frac{\sum_{l \in \text{group } j} \text{Prob}[\mathbf{Z}_{lk} == 1 | \mathbf{x}, \tau = \hat{\tau}]}{\sum_j \sum_{l \in \text{group } j} \text{Prob}[\mathbf{Z}_{lk} == 1 | \mathbf{x}, \tau = \hat{\tau}]}.$$

ω could be fixed at $1/G$ again, and α would be chosen to satisfy equation (11). With this method, observations that are in group j in the initial grouping based on $\hat{\mathbf{Z}}_m$ are most likely to be redistributed into groups close to group j, whereas the Uniform Distance method redistributes the observations uniformly into the groups. Another adaptive variant of Z–distance sampling, ZD–2, is discussed in the text.

References

Atwood, L. D., A. F. Wilson, R. C. Elston, and L. E. Bailey-Wilson (1992). Computational aspects of fitting a mixture of two normal distributions using maximum likelihood. *Communications in Statistics, Part B – Simulation and Computation 21*, 769–781.

Barrett, M. T., P. C. Galipeau, C. A. Sanchez, M. J. Emond, and B. R. Reid (1996). Determination of the frequency of loss of heterozygosity in esophageal adenocarcinoma by cell sorting, whole genome amplification and microsatellite polymorphisms. *Oncogene 12*, 1873–8.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated complete likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 719–725.

Celeux, G. (1997). Discussion of the paper by Richardson and Green. *J Royal Stat Soc B 59*, 775–776.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*, 957–970.

Cheeseman, P. and J. Stutz (1995). Bayesian classification (AutoClass): Theory and results. In U.Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurasamy (Eds.), *Advances in knowledge discovery and data mining*, pp. 153–180. Menlo Park, Calif.: AAAI Press.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chickering, D. M. and D. Heckerman (1996). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. In *Proceedings of the 12th con-*

ference on uncertainty in artificial intelligence, pp. 158–168. Morgan Kaufman Publishers, San Francisco, Calif.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B 39*, 1–22.

Desai, M. (2000). *Mixture Models for Genetic Changes in Cancer Cells*. Ph. D. thesis, University of Washington, Department of Biostatistics.

Desai, M. and M. Emond (2001). A new importance sampling method to compute bayes factors for mixure models with application to allelic-loss data. Technical Report 173, Department of Biostatistics, University of Washington.

Evans, M. and T. Swartz (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems (disc: V11 p54-64). *Statistical Science 10*, 254–272.

Feng, Z. D. and C. E. McCulloch (1996). Using bootstrap likelihood ratios in finite mixture models. *J.R. Statist. Soc B 58*, 609–617.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (2000, October). Model-based clustering, discriminant analysis, and density estimation. Technical Report 380, Department of Statistics, University of Washington.

Geyer, C. J. (1991, December). Reweighting Monte Carlo mixtures. Technical Report 518, School of Statistics, University of Minnesota.

Givens, G. H. and A. E. Raftery (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association 91*, 132–141.

Grunwald, G. K., A. E. Raftery, and P. Guttorp (1993). Time series of continuous proportions. *Journal of the Royal Statistical Society, Series B 55*, 103–116.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics 37*, 185–194.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Clarendon Press.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Keribin, C. (1998). Consistent estimate of the order of mixture models. *Comptes Rendues de l'Academie des Sciences, série I — Mathématiques 326*, 243–248.

Leroux, B. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics 20*, 1350–1360.

Lewis, S. M. and A. E. Raftery (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association 92*, 648–655.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications.* Hayward, Calif.: Institute of Mathematical Statistics.

Neal, R. M. (1998). Erroneous results in "Marginal likelihood from the Gibbs output". http://www.cs.utoronto.ca/˜radford.

Newton, M. A., M. N. Gould, C. A. Reznikoff, and J. D. Haag (1998). On the statistical analysis of allelic-loss data. *Statistics in Medicine 17*, 1425–45.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B 56*, 3–49.

Oh, M.-S. and J. O. Berger (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association 88*, 450–456.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology 25*, 111–193.

Raftery, A. E. (1996a). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika 83*, 251–266.

Raftery, A. E. (1996b). Hypothesis testing and model selection. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall.

Raghavan, N. and D. D. Cox (1998). Adaptive mixture importance sampling. *Journal of Statistical Computation and Simulation 60*, 237–259.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B 59*, 731–792.

Roeder, K. and L. Wasserman (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92*, 894–902.

Rozenkranz, S. L. and A. E. Raftery (1994). Covariate selection in hierarchical models of hospital admission counts: A Bayes factor approach. Technical Report 268, Department of Statistics, University of Washington.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Shibagaki, I., Y. Shimada, T. Wagata, M. Ikenaga, M. Imamura, and K. Ishizaki (1994). Allelotype analysis of esophageal squamous cell carcinoma. *Cancer Res 54*, 2996–3000.

Stephens, M. (1997). Contribution to the discussion of Richardson and Green, 1997: on the Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B 59*, 768–769.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 795–809.

Stephens, M. and P. Donnelly (2000). Inference in molecular population genetics (with discussion). *Journal of the Royal Statistical Society, Series B 62*, 605–635.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B 31*, 80–88.

Wei, G. C. G. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms. *Journal of the American Statistical Association 85*, 699–704.

West, M. (1993). Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society, Series B, Methodological 55*, 409–422.

Zhang, P. (1996). Nonparametric importance sampling. *Journal of the American Statistical Association 91*, 1245–1253.